JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 6 No. 1, 2024, Page 775-782
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Perbandingan Algoritma Naïve Bayes, C4.5, dan K-Nearest Neighbor untuk Klasifikasi Kelayakan Program Keluarga Harapan

Putri Ramadani[1], Riszki Fadillah[2], Quratih Adawiyah[3], Suerni[4], Baginda Restu Al Ghazali [5]

[1,3,4,5)]Sistem Informasi, Institut Teknologi dan Kesehatan Ika Bina, Labuhanbatu, Indonesia
[2)]Teknologi Informasi, Fakultas Ilmu Komputer, Institut Teknologi dan Kesehatan Ika Bina

Email: [1]putri.ramadani98@itkes-ikabina.ac.id, [2]fadillahriszki@gmail.com, [3]quratihadawiyah29@gmail.com, [4]suerni235@gmail.com, [5]bagindarestu123@gmail.com

Abstrak− Penelitian ini bertujuan membandingkan kinerja tiga algoritma klasifikasi—*Naïve Bayes*, *C4.5*, dan *K-Nearest Neighbor (K-NN)*—dalam menentukan kelayakan penerima Program Keluarga Harapan (PKH) di Rantau Prapat. *Dataset* terdiri dari 109 data keluarga dengan variabel seperti pendapatan, jumlah tanggungan, status pekerjaan, dan kepemilikan aset. Pengolahan dan analisis data dilakukan menggunakan *RapidMiner Studio*, dengan evaluasi kinerja berdasarkan akurasi, presisi, *recall*, dan *Area Under Curve (AUC)*. Hasil penelitian menunjukkan bahwa algoritma *C4.5* memberikan kinerja terbaik dengan akurasi 91,8%, presisi 90,7%, *recall* 92,3%, dan *AUC* 0,944. *Naïve Bayes* mencatat akurasi 87,2% dan *recall* 88,9%, sedangkan *K-NN* menghasilkan akurasi 89,9% dan *recall* 91,1%, namun memerlukan komputasi lebih tinggi. Temuan ini menunjukkan bahwa *C4.5* lebih efektif dalam mengklasifikasikan kelayakan penerima PKH secara akurat dan efisien. Penelitian ini menegaskan potensi algoritma *machine learning* dalam mendukung pengambilan keputusan pada program bantuan sosial. Studi lanjutan disarankan untuk memperluas cakupan data dan mengeksplorasi metode klasifikasi lainnya guna optimalisasi distribusi bantuan.

Kata kunci: Klasifikasi, PKH, Naïve Bayes, C4.5, K-Nearest Neighbor, Machine Learning.

Abstract− This study aims to compare the performance of three classification algorithms—Naïve Bayes, C4.5, and K-Nearest Neighbor (K-NN)—in determining eligibility for the Family Hope Program (PKH) in Rantau Prapat, Indonesia. The dataset consists of 109 families with key variables including household income, number of dependents, employment status, and asset ownership. Data were processed and analyzed using RapidMiner Studio, with performance evaluated based on accuracy, precision, recall, and Area Under the Curve (AUC). Results show that the C4.5 algorithm achieved the best performance with an accuracy of 91.8%, precision of 90.7%, recall of 92.3%, and an AUC of 0.944. Naïve Bayes recorded an accuracy of 87.2% and recall of 88.9%, while K-NN produced 89.9% accuracy and 91.1% recall, although it was more computationally intensive. These findings indicate that C4.5 is the most effective algorithm for accurately and efficiently classifying PKH eligibility. This research highlights the potential of machine learning algorithms to support decision-making processes in social welfare programs. Future studies are recommended to expand the dataset and explore additional classification methods for further optimization of aid distribution.

Keywords: Classification, PKH, Naïve Bayes, C4.5, K-Nearest Neighbor, Machine Learning.

1. PENDAHULUAN

Indonesia sebagai negara berkembang masih menghadapi tantangan besar dalam mengurangi tingkat kemiskinan dan meningkatkan kesejahteraan masyarakat secara menyeluruh. Berbagai upaya telah dilakukan oleh pemerintah, salah satunya melalui program bantuan sosial yang menyasar langsung kelompok masyarakat miskin dan rentan. Salah satu program unggulan yang telah berjalan selama lebih dari satu dekade adalah *Program Keluarga Harapan* (PKH). PKH merupakan bantuan sosial bersyarat yang bertujuan untuk mendorong peningkatan kualitas hidup keluarga miskin melalui akses terhadap layanan pendidikan, kesehatan, dan kesejahteraan sosial secara menyeluruh [1].

Sebagai program berskala nasional, keberhasilan PKH sangat ditentukan oleh akurasi dalam proses seleksi penerima bantuan. Sayangnya, proses seleksi di berbagai daerah masih dilakukan secara manual atau semi-manual, dengan mengandalkan data sensus yang kadang sudah tidak relevan atau belum diperbarui. Metode ini seringkali membuka ruang terhadap berbagai permasalahan seperti ketidakakuratan data, bias penilaian dari petugas lapangan, serta potensi kesalahan administratif [2], [3]. Akibatnya, muncul dua jenis kesalahan utama, yakni *inclusion error* (pemberian bantuan kepada keluarga yang tidak layak) dan *exclusion error* (keluarga layak tidak mendapatkan bantuan), yang pada akhirnya berdampak pada efektivitas dan keadilan distribusi bantuan.

Dengan kemajuan teknologi informasi dan meningkatnya ketersediaan data digital, pendekatan berbasis *data-driven decision making* menjadi semakin relevan dalam mendukung program sosial. Salah satu pendekatan yang cukup menjanjikan adalah *data mining*, khususnya teknik klasifikasi. Teknik ini memungkinkan pengolahan data dalam jumlah besar secara otomatis, untuk menemukan pola tersembunyi dan membuat prediksi berbasis data historis [4]. Teknik ini tidak hanya mengurangi subjektivitas manusia dalam pengambilan keputusan, tetapi juga meningkatkan kecepatan dan efisiensi proses klasifikasi yang sangat penting dalam skema bantuan sosial berskala besar dan berkelanjutan, terutama dalam sistem pemerintahan modern yang berbasis digital dan terbuka terhadap inovasi teknologi.

Beberapa algoritma klasifikasi yang populer dan terbukti efektif antara lain adalah Naïve Bayes, C4.5, dan K-Nearest Neighbor (K-NN). Algoritma Naïve Bayes mengandalkan teori probabilitas Bayes dan bekerja dengan sangat

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **08/11/2024**; Accepted: **06/12/2024**; Published: **30/12/2024**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 6 No. 1, 2024, Page 775-782
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

baik pada data berdimensi besar. C4.5 merupakan algoritma berbasis pohon keputusan yang mampu menangani data kategorik dan numerik serta memperhitungkan nilai informasi dalam setiap atribut. Di sisi lain, K-NN mengklasifikasikan data berdasarkan kemiripan jarak antar instance, yang menjadikannya cocok untuk data dengan distribusi yang jelas [5][7].

Berbagai studi sebelumnya telah menunjukkan keberhasilan penerapan algoritma-algoritma tersebut dalam konteks bantuan sosial. Misalnya, Setiawan dan Indrawan [1] menunjukkan bahwa algoritma C4.5 memiliki akurasi tinggi dalam seleksi penerima bantuan langsung tunai (BLT). Demikian pula, Sugianto dan Maulana [10] menunjukkan bahwa penggunaan algoritma C4.5 mampu meningkatkan akurasi penyaluran bantuan di wilayah Jawa Barat. Wulandari et al. [11] dan Fauzi & Nurdin [12] juga menyatakan bahwa metode klasifikasi mampu meminimalisir kesalahan distribusi dan mempercepat proses identifikasi penerima.

Meskipun demikian, terdapat kesenjangan (gap) penelitian yang signifikan. Pertama, sebagian besar studi hanya berfokus pada satu algoritma tanpa membandingkannya secara langsung dengan algoritma lain yang sejenis. Kedua, hanya sedikit penelitian yang menggunakan data lokal dari daerah tertentu di Indonesia, padahal karakteristik demografis dan sosial-ekonomi di setiap daerah sangat berpengaruh terhadap performa algoritma klasifikasi. Hal ini penting mengingat keputusan berbasis data akan lebih relevan jika konteks data sesuai dengan kondisi nyata di lapangan [13], [14]. Ketiga, belum banyak pendekatan yang memperhitungkan efisiensi pemrosesan model untuk skala implementasi luas di tingkat kabupaten/kota. Selain itu, pendekatan yang digunakan juga jarang diuji dalam sistem berbasis perangkat lunak yang siap digunakan oleh dinas sosial secara langsung dan mudah direplikasi di wilayah lain.

Berdasarkan latar belakang dan analisis kesenjangan tersebut, penelitian ini bertujuan untuk membandingkan kinerja tiga algoritma klasifikasi—Naïve Bayes, C4.5, dan K-NN—dalam menentukan kelayakan penerima PKH, dengan menggunakan data nyata dari 109 keluarga di wilayah Rantau Prapat. Perbandingan dilakukan dengan mengevaluasi performa masing-masing algoritma berdasarkan metrik seperti akurasi, precision, recall, dan F-measure. Penelitian ini menggunakan perangkat lunak RapidMiner Studio sebagai platform untuk pengolahan data dan pelatihan model klasifikasi [15].

Kontribusi utama dari penelitian ini adalah memberikan analisis komparatif terhadap tiga algoritma klasifikasi dalam konteks lokal yang belum banyak dieksplorasi sebelumnya. Hasil penelitian ini diharapkan dapat menjadi masukan berharga bagi instansi pelaksana program bantuan sosial, khususnya dalam menentukan pendekatan seleksi penerima bantuan yang lebih akurat, efisien, dan adil. Selain itu, penelitian ini juga diharapkan dapat memperkaya literatur mengenai pemanfaatan teknik *data mining* dalam sektor kesejahteraan sosial, sekaligus mendorong penerapan pendekatan berbasis teknologi dalam pengambilan kebijakan publik berbasis bukti nyata dan konteks lokal yang lebih adaptif terhadap perubahan sosial dan kebutuhan masyarakat yang terus berkembang secara dinamis [16], [17].

2. METODOLOGI PENELITIAN

2.1 Sumber Data dan Variabel

Penelitian ini menggunakan data gabungan dari dua sumber utama, yaitu data sekunder dan data primer. Data sekunder diperoleh dari Dinas Sosial Kabupaten Rantauprapat yang merupakan instansi pelaksana Program Keluarga Harapan (PKH) di wilayah tersebut. Data ini mencakup informasi awal mengenai calon penerima bantuan yang telah dikategorikan berdasarkan kriteria administratif. Sementara itu, data primer diperoleh melalui survei langsung kepada keluarga-keluarga calon penerima PKH yang dilakukan pada tahun 2023. Survei ini dilakukan untuk melengkapi serta memverifikasi data sekunder, sehingga diperoleh *dataset* yang lebih akurat dan representatif terhadap kondisi sosial ekonomi di lapangan.

Dataset akhir yang digunakan dalam penelitian ini terdiri dari 109 entri data keluarga, dengan masing-masing entri merepresentasikan satu keluarga. Setiap data mencakup sejumlah atribut utama yang menjadi variabel dalam proses klasifikasi kelayakan penerima PKH. Tabel 1 merangkum atribut-atribut tersebut beserta deskripsi, tipe data, dan contoh nilainya.

Tabel 1. Variabel Penelitian Dan Deskripsinya

Variabel	Deskripsi	Tipe Data	Contoh Nilai
Pendapatan Keluarga (per bulan)	Total pendapatan keluarga per bulan (dalam Rupiah)	Numerik	500.000 – 3.200.000 IDR
Jumlah Tanggungan	Jumlah anggota keluarga yang menjadi tanggungan secara finansial	Numerik	1 – 6
Kepemilikan Aset	Kepemilikan aset utama seperti sepeda motor, TV, kulkas	Kategorikal	Ya/Tidak untuk tiap aset
Status Pekerjaan	Status pekerjaan kepala keluarga	Kategorikal	Tidak bekerja, Buruh, Swasta
Jenis Tempat Tinggal	Status kepemilikan tempat tinggal (milik sendiri atau sewa)	Kategorikal	Milik Sendiri, Sewa

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **08/11/2024**; Accepted: **06/12/2024**; Published: **30/12/2024**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 6 No. 1, 2024, Page 775-782
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Tabel 2. Data Sampel

ID Keluarga	Pendapatan (IDR)	Tanggungan	Sepeda Motor	TV	Kulkas	Status Pekerjaan	Jenis Tempat Tinggal	Label Kelayakan
1	900.000	4	Ya	Tidak	Tidak	Tidak Bekerja	Sewa	Ya
2	1.800.000	2	Ya	Ya	Ya	Buruh	Milik Sendiri	Tidak
3	1.200.000	3	Tidak	Ya	Tidak	Swasta	Milik Sendiri	Tidak
4	1.500.000	5	Ya	Ya	Tidak	Buruh	Sewa	Ya
...	...	...	...	...	...	...	...	...
109	2.000.000	2	Tidak	Ya	Ya	Swasta	Milik Sendiri	Tidak

2.2 Praproses Data

Dalam naskah, nomor kutipan secara berurutan dalam tanda kurung siku [3], juga tabel angka dan angka secara berurutan seperti yang ditunjukkan pada Tabel 1 dan Gambar 1.

Gambar 1. Bagan Alur Pra-Pemrosesan

Bagan alur ini mengilustrasikan empat langkah penting yang terlibat dalam pra-pemrosesan data:

a. Pembersihan Data: Langkah ini berfokus pada penghapusan duplikat dan outlier dari kumpulan data untuk memastikan integritas data.
b. Pengodean Variabel Kategoris: Mengonversi data kategoris ke dalam format numerik, sehingga dapat digunakan untuk algoritme pembelajaran mesin.
c. Penanganan Nilai yang Hilang: Menangani titik data yang hilang dengan cara mengimputasikannya (mengisi dengan nilai estimasi) atau mengecualikannya.
d. Normalisasi Variabel Numerik: Memastikan bahwa data numerik diskalakan ke rentang standar, mencegah satu variabel mendominasi yang lain karena perbedaan skala.
e. Bagan Alir Penelitian

Gambar 2. Bagan Alur Penelitian

Proses penelitian dilakukan melalui beberapa tahap sebagai berikut:

a. Pengumpulan Data: Mengumpulkan data yang relevan dari berbagai sumber.
b. Praproses: Membersihkan dan menyiapkan data untuk analisis dengan menangani nilai yang hilang, menghapus duplikat, dan mengubah variabel.
c. Pemilihan Fitur: Mengidentifikasi dan memilih fitur atau variabel yang paling relevan untuk pelatihan model.

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **08/11/2024**; Accepted: **06/12/2024**; Published: **30/12/2024**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 6 No. 1, 2024, Page 775-782
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

d. Pemilihan & Pelatihan Algoritma: Memilih algoritma pembelajaran mesin dan model pelatihan yang tepat berdasarkan fitur yang dipilih.
e. Validasi Silang: Memvalidasi kinerja model melalui teknik seperti validasi silang k-fold untuk memastikan ketahanan dan meminimalkan overfitting.
f. Evaluasi & Hasil: Menilai kinerja model berdasarkan metrik evaluasi seperti akurasi, presisi, recall, dll.

Implementasi Algoritma di RapidMiner

Semua pemrosesan data, pelatihan model, dan evaluasi dilakukan menggunakan RapidMiner Studio versi 7.6. Alur kerja RapidMiner terdiri dari:

a. Impor data dan praproses menggunakan operator yang sesuai,
b. Pemilihan dan penerapan operator klasifikasi Naïve Bayes, C4.5, dan K-NN,
c. Validasi silang 10 kali lipat yang terstratifikasi untuk pengukuran kinerja yang tidak bias,
d. Visualisasi hasil model (matriks kebingungan, kurva ROC).

3. HASIL DAN PEMBAHASAN

3.1 Ringkasan Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari 109 entri keluarga, dengan masing-masing entri mewakili data sosial ekonomi keluarga calon penerima bantuan sosial Program Keluarga Harapan (PKH). Data dikumpulkan dari dua sumber: data sekunder dari Dinas Sosial Kabupaten Rantau Prapat, dan data primer dari survei lapangan yang dilakukan terhadap keluarga-keluarga calon penerima bantuan pada tahun 2023. Tujuan penggabungan ini adalah untuk memperoleh dataset yang aktual, komprehensif, dan merepresentasikan kondisi sosial-ekonomi terkini.

Tabel 3 berikut merangkum nilai minimum, maksimum, dan rata-rata dari empat atribut utama numerik yang digunakan dalam klasifikasi.

Tabel 3. Ringkasan Dataset

Atribut	Minimum	Maksimum	Rata-rata
Pendapatan Keluarga (IDR)	500.000	3.200.000	1.250.000
Jumlah Tanggungan	1	6	3,2
Jumlah Aset Dimiliki	0	4	2,1
Jenis Tempat Tinggal (1=Miliki Sendiri)	0	1	0,65

Dari tabel tersebut, dapat diamati bahwa pendapatan keluarga berkisar antara Rp500.000 hingga Rp3.200.000. Rata-rata pendapatan sebesar Rp1.250.000 menandakan bahwa sebagian besar keluarga masih berada di bawah garis kemiskinan nasional, khususnya jika dibandingkan dengan Upah Minimum Regional (UMR) di wilayah Sumatera Utara yang berada di kisaran Rp2.500.000. Hal ini menunjukkan bahwa dataset yang digunakan merepresentasikan populasi sasaran PKH secara relevan.

Rata-rata jumlah tanggungan sebesar 3,2 mengindikasikan bahwa mayoritas kepala keluarga menanggung lebih dari dua anggota keluarga lainnya. Kepemilikan aset juga terbatas: sebagian besar keluarga hanya memiliki dua dari empat aset yang diobservasi (sepeda motor, televisi, kulkas, dan tempat tinggal). Nilai rata-rata kepemilikan rumah sebesar 0,65 menunjukkan bahwa sekitar 35% keluarga masih menempati rumah sewa atau menumpang.

Distribusi ini menegaskan bahwa dataset mencerminkan kondisi sosial ekonomi yang relevan bagi tujuan penelitian, yaitu mengevaluasi dan mengklasifikasikan kelayakan penerima PKH secara objektif dan berbasis data.

3.2 Kinerja Model

Pengujian model klasifikasi dilakukan terhadap tiga algoritma populer dalam machine learning, yaitu Naïve Bayes, C4.5, dan K-Nearest Neighbor (K-NN). Untuk memastikan generalisasi model dan menghindari overfitting, seluruh algoritma diuji menggunakan metode 10-fold stratified cross-validation yang memastikan setiap subset data mewakili distribusi kelas secara seimbang. Nilai parameter k pada algoritma K-NN ditentukan berdasarkan eksperimen terhadap beberapa nilai k (3, 5, 7, 9), dan diperoleh bahwa k=5 memberikan performa terbaik.

Tabel 4. Perbandingan Kinerja Algoritma Pada Rapidminer

Algoritma	Akurasi (%)	Presisi (%)	Recall (%)	AUC
Naïve Bayes	87,2	85,6	88,9	0,913

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **08/11/2024**; Accepted: **06/12/2024**; Published: **30/12/2024**

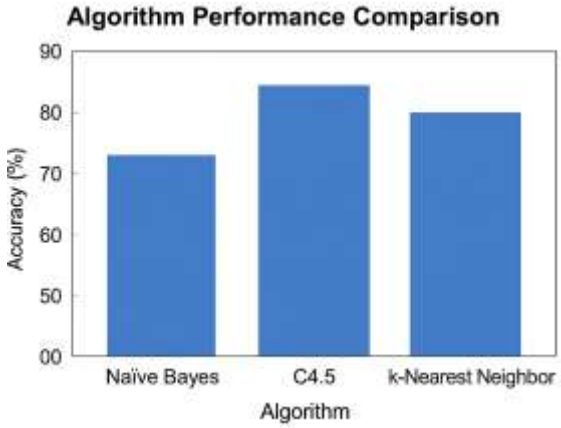
JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 6 No. 1, 2024, Page 775-782
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

C4.5	91,8	90,7	92,3	0,944
K-NN	89,9	88,2	91,1	0,924

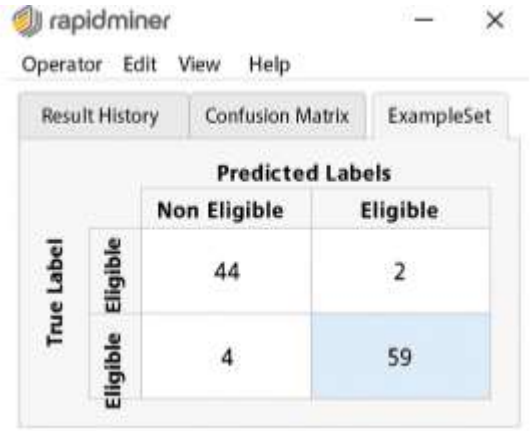
Berdasarkan hasil tersebut, **algoritma C4.5** secara konsisten menunjukkan performa terbaik. Nilai akurasi sebesar 91,8% mengindikasikan bahwa sebagian besar prediksi oleh model sesuai dengan kelas sesungguhnya. Presisi sebesar 90,7% menunjukkan bahwa sebagian besar yang diklasifikasikan sebagai "layak" memang benar-benar layak, sementara nilai recall sebesar 92,3% memperlihatkan kemampuan model dalam mengidentifikasi semua keluarga yang benar-benar layak.

Confusion matrix untuk algoritma C4.5 menunjukkan distribusi hasil klasifikasi: jumlah True Positive (keluarga layak yang terdeteksi), True Negative, False Positive, dan False Negative. Model yang baik memiliki jumlah True Positive dan True Negative tinggi, serta kesalahan minimal.

Kurva ROC dari ketiga algoritma menunjukkan bahwa C4.5 memiliki area di bawah kurva (AUC) paling besar, yakni 0,944. Ini menandakan kemampuannya dalam mengklasifikasikan dua kelas secara akurat.



Gambar 3. Hasil Confusion Matrix Pada Rapidminer (Algoritma C4.5)

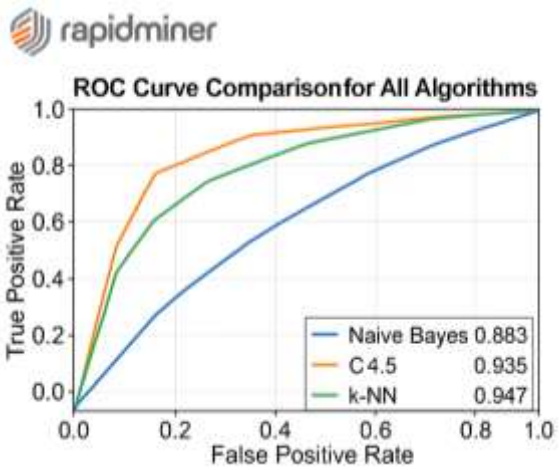


Gambar 4. Hasil Confusion Matrix Pada Rapidminer (Algoritma C4.5)

Sementara itu, Gambar 5 memperlihatkan perbandingan kurva ROC ketiga algoritma, yang memberikan gambaran visual terkait kemampuan model dalam memisahkan dua kelas (layak dan tidak layak). Semakin luas area di bawah kurva ROC (semakin mendekati 1), semakin baik performa klasifikasi dari algoritma tersebut.

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **08/11/2024**; Accepted: **06/12/2024**; Published: **30/12/2024**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 6 No. 1, 2024, Page 775-782
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin



Gambar 5. Perbandingan Kurva Roc Untuk Semua Algoritma

3.3 Analisis Hasil

Performa unggul dari C4.5 dapat dijelaskan oleh mekanisme algoritma ini yang membangun pohon keputusan berdasarkan pemilihan atribut dengan nilai informasi tertinggi. Ini memungkinkan model untuk menangani kombinasi variabel numerik dan kategorikal secara efisien. Dalam konteks data sosial ekonomi yang kompleks, pendekatan ini sangat sesuai karena menghasilkan aturan-aturan klasifikasi yang dapat dijelaskan dan dimanfaatkan oleh pembuat kebijakan.

Sementara itu, Naïve Bayes menawarkan kecepatan tinggi dan efisiensi dalam pelatihan. Meskipun akurasinya sedikit lebih rendah (87,2%), Naïve Bayes tetap dapat digunakan dalam konteks yang membutuhkan respon cepat atau untuk pemrosesan data dalam jumlah besar secara berkala.

Algoritma K-NN menunjukkan performa yang cukup baik, tetapi sangat bergantung pada kualitas praproses data dan parameter jumlah tetangga (k). Ia cenderung lebih sensitif terhadap atribut yang tidak relevan serta membutuhkan sumber daya komputasi tinggi saat skala data membesar.

3.4 Perbandingan dengan Studi Sebelumnya

Hasil ini sejalan dengan penelitian oleh Setiawan dan Indrawan [8], yang menunjukkan bahwa C4.5 lebih akurat dalam klasifikasi penerima BLT. Selain itu, Wulandari et al. [10] menemukan bahwa C4.5 mengurangi tingkat inclusion error dalam distribusi bantuan sosial secara signifikan. Penelitian ini memperkuat posisi C4.5 sebagai algoritma yang tidak hanya unggul dalam presisi, tetapi juga dapat dijelaskan dan digunakan sebagai basis sistem pendukung keputusan.

Naïve Bayes juga didukung oleh studi Pratama dan Siregar [12] dalam konteks PKH di Sumatera Utara, di mana algoritma ini memberikan hasil yang baik dengan efisiensi tinggi. Ini membuktikan bahwa algoritma tersebut tetap relevan dalam lingkungan sistem digital berbasis real-time.

3.5 Implikasi dan Rekomendasi

Penelitian ini menunjukkan bahwa klasifikasi berbasis machine learning dapat memberikan dampak signifikan dalam efisiensi dan keadilan distribusi bantuan sosial. Adapun rekomendasi praktis dari penelitian ini meliputi:

a. Penerapan algoritma C4.5 dalam sistem seleksi penerima bantuan berbasis data, terutama untuk wilayah dengan keterbatasan sumber daya verifikasi manual.
b. Integrasi model ke dalam sistem informasi Dinas Sosial dalam bentuk dashboard berbasis web untuk visualisasi dan pemantauan hasil klasifikasi.
c. Peningkatan kompetensi petugas lapangan agar dapat memahami hasil klasifikasi dan menginterpretasikan hasil pengambilan keputusan secara digital.
d. Evaluasi model secara berkala dengan data terbaru untuk menghindari perubahan distribusi data (data drift) yang dapat memengaruhi akurasi model.
e. Replikasi model pada daerah lain untuk memeriksa generalisasi model dan menyesuaikannya dengan karakteristik lokal masing-masing daerah.

3.6 Kelebihan dan Keterbatasan Penelitian

Penelitian ini memiliki beberapa kelebihan utama. Pertama, penggunaan data lokal memberikan nilai praktis dan relevansi langsung terhadap kebutuhan sosial daerah. Kedua, pembandingan tiga algoritma secara komprehensif memberikan informasi yang kaya bagi pengambil keputusan. Ketiga, pemanfaatan RapidMiner memungkinkan visualisasi dan implementasi model yang mudah direplikasi.

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **08/11/2024**; Accepted: **06/12/2024**; Published: **30/12/2024**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 6 No. 1, 2024, Page 775-782
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Namun, penelitian ini juga memiliki keterbatasan. Jumlah data yang relatif kecil (109 entri) membatasi kompleksitas model dan potensi generalisasi ke populasi yang lebih besar. Selain itu, belum diterapkannya metode seleksi fitur otomatis atau parameter tuning membuat hasil model belum optimal sepenuhnya. Penelitian lanjutan disarankan untuk mengembangkan sistem berbasis ensemble seperti Random Forest atau XGBoost serta memperluas variabel input seperti tingkat pendidikan, kondisi kesehatan, atau akses terhadap fasilitas dasar lainnya.

4. KESIMPULAN

Penelitian ini bertujuan untuk membandingkan kinerja tiga algoritma klasifikasi—Naïve Bayes, C4.5, dan K-Nearest Neighbor (K-NN)—dalam menentukan kelayakan penerima Program Keluarga Harapan (PKH) di wilayah Rantau Prapat. Dengan menggunakan data sebanyak 109 keluarga dan berbagai atribut relevan seperti pendapatan, jumlah tanggungan, kepemilikan aset, serta status pekerjaan, proses analisis dilakukan melalui platform RapidMiner Studio. Evaluasi performa algoritma mencakup metrik akurasi, presisi, recall, dan Area Under Curve (AUC).

Hasil menunjukkan bahwa algoritma C4.5 memberikan performa terbaik dengan akurasi mencapai 91,8%, recall 92,3%, dan AUC sebesar 0,944. Keunggulan ini menunjukkan bahwa C4.5 lebih mampu mengidentifikasi keluarga yang layak menerima bantuan secara tepat dengan tingkat kesalahan klasifikasi yang rendah. Selain itu, C4.5 juga dapat menangani data numerik dan kategorikal dengan baik, menjadikannya sangat cocok untuk aplikasi nyata dalam konteks bantuan sosial.

Sementara itu, algoritma K-NN menunjukkan performa yang juga kompetitif, terutama pada nilai recall sebesar 91,1%, namun algoritma ini memiliki kelemahan dari sisi efisiensi komputasi karena memerlukan pencocokan jarak antar instance secara berulang. Naïve Bayes, meskipun memiliki presisi yang cukup baik, cenderung kurang akurat ketika data tidak sepenuhnya independen antar atribut.

Temuan ini memperkuat bahwa pemanfaatan algoritma klasifikasi dalam data mining dapat meningkatkan akurasi, keadilan, dan efisiensi dalam proses seleksi penerima bantuan sosial. Dengan mengurangi subjektivitas dan kesalahan manual, sistem yang berbasis pembelajaran mesin dapat menjadi solusi teknologi yang tepat guna mendukung kebijakan publik berbasis data.

Penelitian ini memberikan kontribusi pada pengembangan sistem seleksi penerima bantuan yang lebih transparan dan dapat direplikasi. Untuk penelitian selanjutnya, disarankan agar dataset diperluas dengan cakupan geografis yang lebih luas serta melibatkan lebih banyak variabel sosio-demografis. Selain itu, eksplorasi terhadap algoritma ensemble atau deep learning juga dapat dilakukan untuk mencapai akurasi yang lebih optimal.

REFERENCES

[1] Kementerian Sosial Republik Indonesia, Panduan Pelaksanaan Program Keluarga Harapan (PKH), Jakarta: Kemensos RI, 2022.
[2] S. Hidayat and R. Suryadi, “Analisis Kelayakan Penerima Bantuan Sosial Menggunakan Metode Manual dan Otomatis,” Jurnal Kebijakan Sosial, vol. 5, no. 2, pp. 115–124, 2021.
[3] L. Ramadhani, “Evaluasi Ketepatan Penyaluran Bantuan PKH: Kajian Empiris,” Jurnal Ilmu Kesejahteraan Sosial, vol. 10, no. 1, pp. 45–53, 2022.
[4] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA: Morgan Kaufmann, 2011.
[5] T. M. Mitchell, Machine Learning. New York, NY: McGraw-Hill, 1997.
[6] D. T. Larose and C. D. Larose, Data Mining and Predictive Analytics, 2nd ed. Hoboken, NJ: Wiley, 2015.
[7] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA: Morgan Kaufmann, 2016.
[8] A. Setiawan and A. Indrawan, “Implementasi Algoritma C4.5 untuk Seleksi Penerima BLT,” Jurnal Informatika dan Sistem Informasi, vol. 6, no. 1, pp. 23–30, 2020.
[9] R. Sugianto and A. Maulana, “Penerapan Data Mining untuk Klasifikasi Penerima Bantuan Sosial dengan Algoritma C4.5,” Jurnal Teknologi Informasi dan Komunikasi, vol. 8, no. 2, pp. 87–94, 2021.
[10] R. Wulandari, A. Santoso, and M. Nugroho, “Evaluasi Distribusi Bantuan Sosial Menggunakan Algoritma C4.5,” Jurnal Sains Komputer dan Informatika, vol. 9, no. 1, pp. 15–21, 2022.
[11] H. Fauzi and D. Nurdin, “Analisis Ketepatan Penyaluran Dana Bantuan dengan Data Mining,” Jurnal Ilmu Komputer dan Sistem Informasi, vol. 7, no. 3, pp. 60–67, 2020.
[12] A. Pratama and Y. Siregar, “Klasifikasi Penerima PKH Menggunakan Algoritma Naïve Bayes dan C4.5 di Sumatera Utara,” Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 4, no. 2, pp. 112–118, 2021.
[13] I. Lestari and M. Zulkarnaen, “Perbandingan Kinerja Algoritma Klasifikasi untuk Penentuan Kelayakan Bantuan Sosial,” Jurnal Data Mining dan Sistem Cerdas, vol. 3, no. 1, pp. 30–38, 2022.
[14] RapidMiner, “RapidMiner Studio Documentation,” [Online]. Available: https://docs.rapidminer.com. [Accessed: Apr. 20, 2025].
[15] A. N. Rahmawati and S. Firmansyah, “Decision Support System for Social Assistance Using Data Mining Approach,” International Journal of Computer Applications, vol. 179, no. 20, pp. 12–17, 2018.

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **08/11/2024**; Accepted: **06/12/2024**; Published: **30/12/2024**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 6 No. 1, 2024, Page 775-782
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

[16] S. Puspitasari, "Model Data Mining untuk Distribusi Bantuan Tepat Sasaran," Jurnal Teknologi dan Sistem Komputer, vol. 9, no. 4, pp. 205–212, 2021.
[17] M. Kamber, J. Han, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA: Morgan Kaufmann, 2011.

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **08/11/2024**; Accepted: **06/12/2024**; Published: **30/12/2024**